# TrendLearner: Early Prediction of Popularity Trends of User Generated Content

Flavio Figueiredo[1,*], Jussara M. Almeida[a], Marcos A. Gonçalves[a], Fabricio Benevenuto[a]

[a]*Department of Computer Science, Universidade Federal de Minas Gerais*
*Av. Antônio Carlos 6627, CEP 31270-010, Belo Horizonte - MG, Brazil. Phone: +55 (31) 3409-7541, Fax: +55 (31) 3409-5858*



**Abstract**

Predicting the popularity of user generated content (UGC) is a valuable task to content providers, advertisers, as well as social media researchers. However, it is also a challenging task due to the plethora of factors that affect content popularity in social systems. Here, we focus on the problem of predicting the popularity *trend* of a piece of UGC (object) *as early as possible*. Unlike previous work, we explicitly address the inherent tradeoff between prediction accuracy and remaining interest in the object after prediction, since, to be useful, accurate predictions should be made *before* interest has exhausted. Given the heterogeneity in popularity dynamics across objects, this tradeoff has to be solved on a per-object basis, making the prediction task harder. We tackle this problem with a novel two-step learning approach in which we: (1) extract popularity trends from previously uploaded objects, and then (2) predict trends for newly uploaded content. Our results for YouTube datasets show that our classification effectiveness, captured by F1 scores, is 38% better than the baseline approaches. Moreover, we achieve these results with up to 68% of the views still remaining for 50% or 21% of the videos, depending on the dataset.

*Keywords:* popularity, trends, classification, social media, ugc, prediction

## 1. Introduction

The success of Internet applications based on user generated content (UGC)[1] has motivated questions such as: How does content popularity evolve over time? What is the potential popularity a piece of content will achieve after a given time period? How can we predict popularity evolution of a particular piece of UGC? For example, from a system perspective, accurate popularity predictions can be exploited to build more cost-effective content organization and delivery platforms (e.g., caching systems, CDNs). They can also drive the design of better analytic tools, a major segment nowadays [20, 34], while online advertisers may benefit from them to more effectively place contextual advertisements. From a social perspective, understanding issues related to popularity prediction can be used to better understand the human dynamics of consumption. Moreover, being able to predict popularity on an automated way is crucial for marketing campaigns (e.g. created by activists or politicians), which increasingly often use the Web to influence public opinion.

**Challenges:** However, predicting the popularity of a piece of content, here referred to as an *object*, in a social system is a very challenging task. This is mostly due to the various phenomena affecting the popularity prediction of social media – which were observed on the datasets we use (as well as others) [11, 22, 33] – as well as the diminishing interesting in objects over time, which implies that popularity predictions must be timely to capture user interest and be useful in real work settings. Both challenges can be summarized as follows:

1. Due to the easiness with which UGC can be created, many factors can affect an object's popularity. Such factors include, for instance, the object's content, the social context in which it is inserted (e.g., social neighborhood or influence zone of the object's creator), the mechanisms used to access the content (e.g., searching, recommendation, toplists), or even an external factor, such as a hyperlink to the content in a popular blog or website. These factors can cause spikes in the surge of interest in objects, as well as information propagation cascades which affect the popularity trends of objects.
2. To be useful in a real scenario, a popularity prediction approach must identify popularity trends *before the user interest in the object has severely diminished.* To illustrate this point, Figure 1 shows the popularity evolution of two YouTube videos: the video on the left receives more than 80% (shaded region) of all views received during its lifespan in the first 300 days since upload, whereas the other video receives only about half of its total views in the same time frame. If we were to monitor each video for 300 days, most potential views of the first video would be lost. In other words, not all objects require the same monitoring period, as assumed by previous work, to produce accurate predictions: for some objects, the prediction can be made earlier. Thus, the tradeoff should be solved on a *per-object* basis, which implies that determining the duration of the monitoring period that leads to a good solution of the tradeoff for *each object* is part of the problem.

*Corresponding author
*Email addresses:* flaviov@dcc.ufmg.br (Flavio Figueiredo), jussara@dcc.ufmg.br (Jussara M. Almeida), mgoncalv@dcc.ufmg.br (Marcos A. Gonçalves), fabricio@dcc.ufmg.br (Fabricio Benevenuto)

[1] YouTube, Flickr, Twitter, and so forth